Het testen van cognitieve vaardigheden van allochtone leerlingen

van de Vijver, F.J.R.; Willemse, G.R.C.M.; van de Rijt, B.A.M.

Published in:
De Psycholoog

Publication date:
1993

Fons van de Vijver
Godelief Willemse
Bernadette van de Rijt

Het testen van cognitieve vaardigheden van allochtone leerlingen

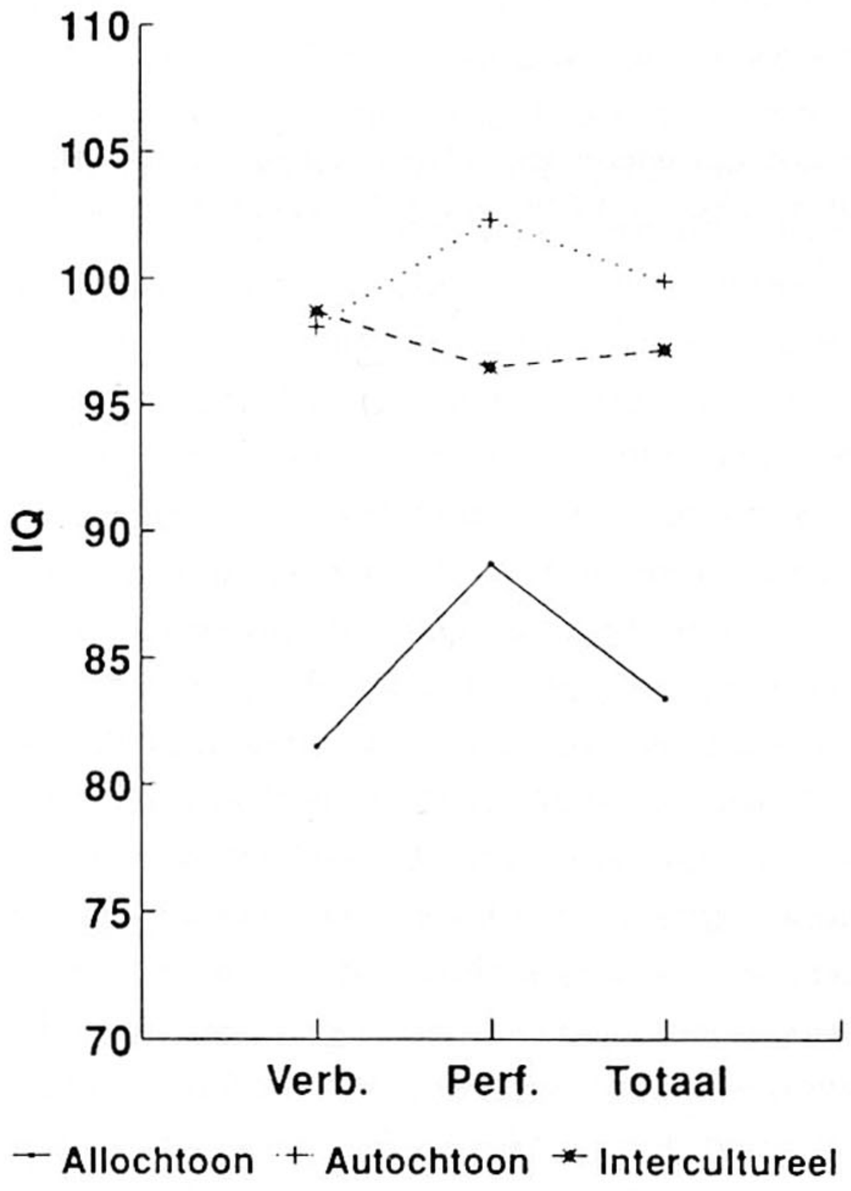
Het aantal allochtone leerlingen in het basisonderwijs is gedurende het laatste decennium sterk gestegen. De auteurs van onderstaand artikel verklaren dat veel bestaande psychologische tests voor deze leerlingen niet goed bruikbaar zijn. Welke alternatieven zijn er en hoe bruikbaar zijn deze?

Nederland is gedurende de laatste decennia een multiculturele samenleving geworden. Het percentage allochtone leerlingen in het onderwijs is sterk gestegen. Zo is volgens het Centraal Bureau voor de Statistiek in het basisonderwijs alleen al in de periode van 1979 tot 1986 dit aantal toegenomen van 5.0% naar 9.6% (cf. Roelandt & Veenman, 1988). Dit heeft tot aanpassingen geleid binnen het onderwijs, zoals de invoering van Internationale Schakelklassen en van een vak als Onderwijs in Eigen Taal en Cultuur.

In dit artikel wordt getracht een overzicht te geven van de huidige stand van zaken betreffende één psychologisch aspect van dit cultureel pluralisme, namelijk het gebruik van tests bij het vaststellen van cognitieve vaardigheden van allochtone leerlingen. Eerst zal een kort overzicht gegeven worden van de bevindingen met conventionele paper-and-pencil tests. Vervolgens komen de problemen met het gebruik van reguliere tests voor cognitieve vaardigheden en schoolvorderingen ter sprake. Voor deze problemen zal een aantal 'remedies' worden besproken. Een evaluatie van de bruikbaarheid van deze remedies vormt het onderwerp van het laatste deel van het artikel.

Conventionele tests

In onderzoek naar cognitieve vaardigheden wordt veel met intelligentietests gewerkt. Als de gemiddelde scores van allochtone en autochtone leerlingen op deze tests vergeleken worden, valt een consistent beeld waar te nemen: de groep allochtone leerlingen vertoont steeds een lager gemiddelde (bijvoorbeeld De Jong, 1987; Resing, Bleichrodt & Drenth, 1986). Ook in eigen onderzoek zien we dit. In figuur 1 zijn de gemiddelde scores op de WISC-R, een voor de Nederlandse populatie genormeerde intelligentietest, weergegeven van 58 allochtone en 49 autochtone leerlingen uit de hoogste groep van het basisonderwijs (Van de Rijt, 1990). Met name op het verbaal IQ vertoonden de allochtone leerlingen een aanzienlijk lager gemid-

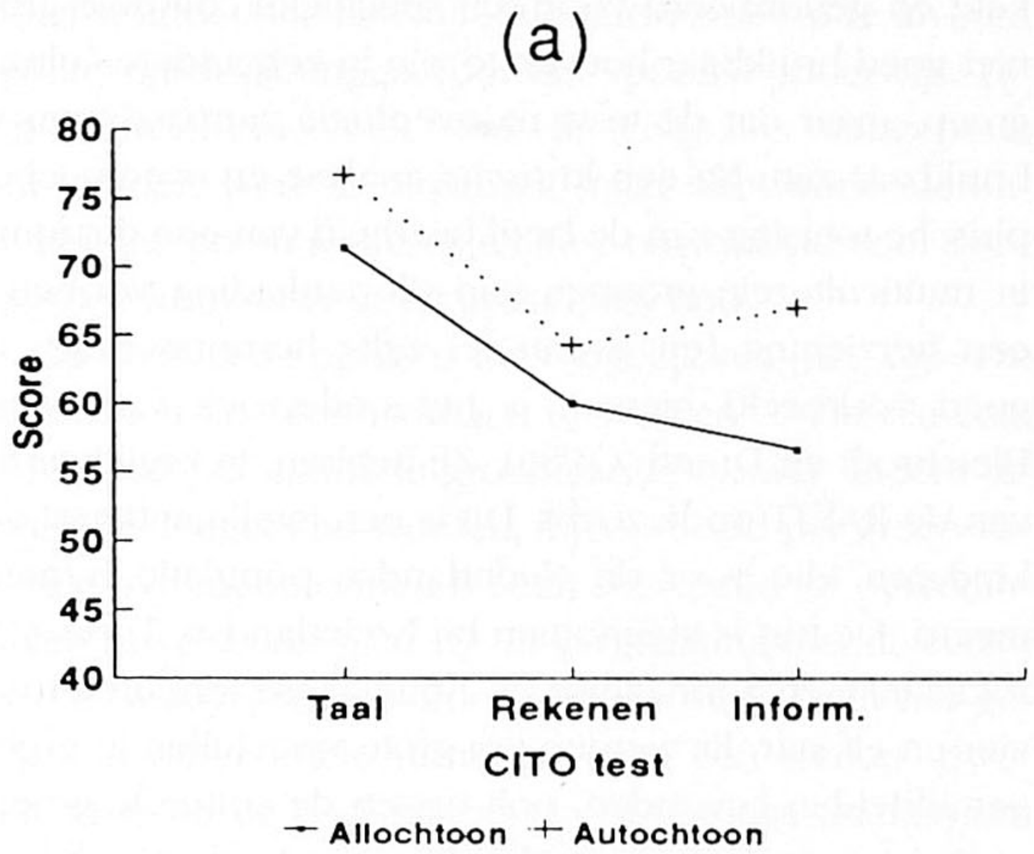


Figuur 1. Gemiddelde WISC-R scores per culturele groep

delde dan de autochtone leerlingen. Verblijfsduur in Nederland speelt hierbij een belangrijke rol. 'Eerste generatie' leerlingen hadden een lagere score dan 'tweede generatie' leerlingen; de gemiddelden van het verbale IQ waren respectievelijk 77.1 en 86.9, van het performale IQ waren deze 83.1 en 95.7. Interessant zijn verder de bevindingen in de 'interculturele' groep; dit betrof 21 kinderen met één autochtone en één allochtone ouder. De resultaten van deze leerlingen en van de autochtone leerlingen waren vrijwel identiek. In cultureel en linguïstisch opzicht lijkt deze interculturele groep dicht bij de autochtone groep te staan.

Dat het taalkundig aspect een belangrijke rol speelt in de prestatieverschillen tussen allochtone en autochtone

leerlingen valt ook af te leiden uit andere tests die in het onderzoek werden afgenomen. Op de Otis, een test voor verbale intelligentie (Maussen, 1988), was het verschil tussen autochtonen en allochtonen naar verhouding groter dan op de Raven, een test voor algemene intelligentie waarop de verbale component veel minder van belang is.



Figuur 2. CITO-Entree scores per culturele groep: (a) Willemse (1989) en (b) Van de Rijt (1990)

Schoolvorderingentests laten eveneens een consistent beeld zien (cf. De Jong, 1987; Van Esch, 1983). In Figuur 2 zijn resultaten weergegeven van de CITO-Entree toets; dit is een kennistoets met drie onderdelen: taal, rekenen en informatie. Deze laatste bevat vragen naar algemene ontwikkeling zoals "Waar kun je waarschijnlijk iets vinden over het klimaat in Griekenland? (a) in een radio- en televisiegids, (b) in een encyclopedie, (c) in een krant, (d) in een treingids?". Zowel Willemse (1989; Van de Vijver & Willemse, 1991) (cf. Figuur 2a) als Van de Rijt (1990) (cf. Figuur 2b) vonden aanzienlijke verschillen in prestaties van allochtone en autochtone leerlingen. In beide onderzoekingen vertoonde de rekentest de kleinste en de informatietest de grootste verschillen.

De vraag rijst nu hoe de gevonden testscoreverschillen tussen allochtone en autochtone leerlingen dienen te worden geïnterpreteerd. Het lijkt onwaarschijnlijk dat de grote verschillen in bijvoorbeeld IQ-scores verwijzen naar even grote verschillen in intelligentie tussen de culturele groepen, hoe consistent deze intergroepsverschillen ook worden gevonden. Het is de vraag of de gevonden verschillen in prestatie niet meer verwijzen naar eigenschappen van het meetinstrument dan van de culturele groepen. In zijn algemeenheid kan worden betwijfeld of conventionele tests valide uitspraken over vergelijking tussen cognitieve vaardigheden van verschillende groepen toestaan. Daarvoor vertonen deze tests te veel problemen.

In de eerste plaats doen conventionele tests een *groot beroep op de verbale vermogens* van de leerling(e). Zowel in de testinstructie als in de formulering van de items kunnen woorden of specifiek Nederlands idioom zijn gebruikt die onbedoeld kapitaliseren op individuele verschillen in verbale vermogens. Met name als verbale vermogens niet zelf getest worden, zal dit onwenselijk zijn, bijvoorbeeld in rekentoetsen.

Vervolgens kunnen – veelal onbedoeld – *culturele waarden en normen* in tests binnensluipen. Soms wordt ook een grote kennis van de Nederlandse cultuur verondersteld. Zo wordt in de WISC-R gevraagd wat spek is. Voor mediterrane leerlingen bleek dit een relatief moeilijk item te zijn (Van de Rijt, 1990). Dit is niet verwonderlijk als men bedenkt dat deze leerlingen een sterk op de Islam georiënteerde culturele achtergrond hebben waarin een voedseltaboe op varkensvlees rust. "Wat is spek?" is naast een toets op woordkennis ook een graadmeter voor de bekendheid met de Nederlandse samenleving. Het probleem dat tests een impliciete meting zijn van de assimilatiegraad van de onderzochte aan de Nederlandse maatschappij, is niet beperkt tot de WISC-R. De meeste tests die we gebruiken, zijn vanuit een Nederlands patroon van normen en waarden samengesteld (Hofstee, 1990). De relatief grote verschillen op de informatietest tussen allochtonen en autochtonen die door Willemse (1989) en Van de Rijt (1990) werden gevonden (cf. Figuur 2) vallen wellicht ook te verklaren uit de grote nadruk op de kennis van de Nederlandse taal en cultuur in deze test.

Tenslotte kan *testwaardigheid* ('test-wiseness', Sarnacki, 1979) een differentiële invloed hebben op de prestaties van culturele groepen. Nederlandse en in Nederland opgeleide leerlingen zullen in het algemeen veel ervaring hebben in het omgaan met tests. In sommige tests wordt gewerkt met een tijdlimiet; de leerling(e) moet bijvoorbeeld gedurende een minuut zoveel mogelijk woorden hardop lezen. Het probleem hierbij is een optimale combinatie te vinden van snelheid en nauwkeurigheid. Eerdere ervaringen met soortgelijke situaties kunnen daarbij behulpzaam zijn. Reguliere toetsen in het onderwijs zullen vaak een goede voorbereiding vormen op psychologische tests.

De storende invloed van de factoren die in de vorige sectie besproken zijn, zal niet voor alle soorten tests even groot zijn (Van de Vijver & Poortinga, 1992). Een belangrijke vraag daarbij is wat met een testscore wordt beoogd. In het onderwijs zijn drie soorten testgebruik te onderscheiden. Soms worden tests afgenomen om inzicht te krijgen in de capaciteiten van een leerling(e); intelligentietests

worden meestal op deze manier gebruikt. De resultaten van deze tests kunnen door bovengenoemde factoren worden verstoord. Met name als met een test gepoogd wordt inzicht te krijgen in het maximum haalbare niveau van een leerling(e), eventueel na extra cursussen en begeleiding, is het van belang de invloed van factoren als taalkennis en testvaardigheid zoveel mogelijk te beperken.

Met schoolvorderingstests zoals de CITO-Entree en -Eindtoets wordt geprobeerd een inzicht te krijgen in de hoeveelheid kennis die een leerling(e) heeft opgebouwd tijdens een curriculum. Als de vragen van de test een redelijke weergave vormen van de inhoud van het curriculum – en niet onbedoelde moeilijkheden bevatten zoals een ingewikkeld geformuleerde vraagstelling, is de kans op storende factoren niet zo groot. De specifieke culturele inhoud van de items kan juist een afspiegeling van het curriculum zijn. De vraag hoeveel kilometer Amsterdam en Haarlem van elkaar liggen kan in een gewone intelligentietest die in een multiculturele context afgenomen wordt, een slecht item zijn terwijl het item aan het eind van een cursus over de ligging van Nederlandse steden accuraat kan zijn.

Een derde testtoepassing doet zich voor bij selectie en plaatsing. Met testscores wordt dan, meestal met een regressievergelijking, een criterium voorspeld; in het onderwijs gaat het veelal om het voorspellen van schoolsucces. Vanuit psychologisch oogpunt is schoolsucces een complexe variabele, waarin onder andere cognitieve, motivationele en sociale elementen een rol spelen. Het is in principe mogelijk dat voor autochtone en allochtone leerlingen niet dezelfde relatie (regressievergelijking) geldt tussen predictor (testscore) en criterium (schoolsucces). Daarnaast kan de vraag gesteld worden of het criterium zelf wel vergelijkbaar is over culturele groepen. Zijn maten voor schoolsucces zoals examenresultaten of docentbeoordelingen een adequate graadmeter van schoolsucces van zowel autochtonen als allochtonen? Veel maten voor schoolsucces zoals examencijfers of rapportpunten komen, direct of indirect, tot stand op basis van resultaten op toetsen die veel lijken op psychologische tests.

Het is dan ook aannemelijk dat de toetsen (en derhalve de schoolsuccesmaten) dezelfde problemen zullen vertonen als conventionele psychologische tests. De bias is dan niet tot de predictor beperkt, maar ook het criterium is niet helemaal vergelijkbaar over culturele groepen. Of in zo'n situatie één regressiefunctie berekend moet worden voor autochtonen en allochtonen wordt nogal eens betwijfeld. Vanuit theoretisch oogpunt is het weinig zinvol om dezelfde regressiefunctie te gebruiken als bias aanwezig is in zowel de predictor als in het criterium. Vanuit praktisch oogpunt valt op te merken dat dit nog niet wil zeggen dat verschillende regressiefuncties zouden gelden voor beide groepen. Het is een empirische vraag of het gebruik van één regressiefunctie tot een systematische overschatting of onderschatting van de prestaties van enige culturele groep leidt.

Remedies

Met nogal uiteenlopende benaderingen is geprobeerd om de bruikbaarheid van tests voor allochtonen te verhogen. We zullen hier vier soorten onderscheiden. In de eerste plaats is het mogelijk om *bestaande instrumenten aan te passen*. Het idee achter deze aanpak is dat tests ontwikkeld en genormeerd voor een specifieke culturele groep, niet goed bruikbaar hoeven te zijn in een andere culturele groep, maar dat de tests na eventuele aanpassingen wel bruikbaar zijn. Na een kritische analyse en eventueel empirische toetsing van de bruikbaarheid van een instrument in multiculturele groepen kan dit aanleiding vormen tot een herziening (en eventueel zelfs hernormering). Een goed voorbeeld hiervan is het onderzoek van Resing, Bleichrodt en Drenth (1986). Zij hebben de bruikbaarheid van de RAKIT onderzocht. Dit is een intelligentietest voor kinderen, die voor de Nederlandse populatie is genormeerd. De test is afgenomen bij Nederlandse, Turkse, Marokkaanse en Surinaamse en Antilliaanse kinderen tussen vier en elf jaar. Er werden vrij grote verschillen in groepsgemiddelden gevonden, ook tussen de culturele groepen onderling: de Turkse en Marokkaanse leerlingen haalden gemiddeld lagere scores dan de Surinaams/Antilliaanse.

De verbale subtests vertoonden de grootste verschillen. Verder bleken ook leeftijd, socio-economische status en verblijfsduur van invloed op de scores. Per culturele groep werd vervolgens een factoranalyse uitgevoerd. Er was een grote overeenstemming tussen de factoren in elk van de groepen, een gangbare bevinding in crosscultureel onderzoek met gestandaardiseerde intelligentietests (cf. Irvine, 1979). Hierbij maakten ze wel de kanttekening dat kennis van de Nederlandse taal niet expliciet onderzocht is bij de kinderen. Hoewel gepoogd was de testinstructies zo helder mogelijk te maken, kon niet worden uitgesloten dat prestatieverschillen deels op differentiële taalkennis terug te voeren zijn. De auteurs adviseren om de subtest 'Woordenschat' bij kinderen die nog niet lang in Nederland zijn niet in de berekening van het IQ te betrekken maar te gebruiken als maat voor beheersing van de Nederlandse taal. Een empirisch onderzoek zoals uitgevoerd door Resing, Bleichrodt en Drenth geeft waardevolle informatie over de mogelijkheden en beperkingen van een test in een multiculturele context.

Een ander (recent) voorbeeld van aanpassing van bestaande tests is te vinden in het onderzoek van Resing (1990). In dit onderzoek zijn bij twee subtests van de RAKIT ('Exclusie' en 'Analogieën', twee inductieve redeneertaken) zogenaamde leerpotentieelprocedures ontwikkeld. Het doel hiervan was na te gaan of maten voor leerpotentieel een aanvulling kunnen vormen op de traditionele intelligentietests. Leerpotentieel wordt hierbij gedefinieerd als "de mate waarin of de efficiëntie waarmee een individu in staat is te profiteren van instructie" (Resing, 1990, p. 23). Als maat voor leerpotentieel is de hoeveelheid hulp genomen die een leerling(e) nodig heeft om een bepaald criterium te bereiken. Het onderzoek is uitgevoerd bij autochtone leerlingen in de leeftijd van 7 tot

8 jaar, afkomstig van basis-, LOM- en MLK-scholen. Uit de resultaten kan onder andere worden afgeleid dat de leerpotentieelscores een substantiële extra bijdrage leveren aan de predictie van schoolprestaties – hierbij dient opgemerkt te worden dat een toename in predictieve validiteit ook het geval is bij herhaalde testafname waarbij geen training is gegeven. Daarnaast blijken maten voor leerpotentieel waardevolle aanvullende informatie op te leveren bij plaatsingsbeslissingen voor het speciaal onderwijs. Dit leerpotentieel-onderzoek zou in enigszins aangepaste vorm kunnen worden uitgebreid naar allochtone leerlingen. Een leerpotentieeltest specifiek ontwikkeld voor deze doelgroep komt later in dit artikel aan bod.

Op de tweede plaats is het mogelijk om met *differentiële normen en beoordelingen* te werken. Deze remedie houdt in dat per culturele groep aan testscores andere interpretaties toegekend worden; bijvoorbeeld per groep variërende minimumnormen bij een sollicitatie, of omzettingen van ruwe scores naar IQ. In vergelijking met de eerste remedie, waarbij tests worden aangepast, zal aan het gebruik van differentiële normen veelal een sterker afwijzende visie op de bruikbaarheid van het instrumentarium ten grondslag liggen. Differentiële normen worden dan geïntroduceerd om te corrigeren, voor al dan niet terecht veronderstelde bias, zoals ongelijke groepsgemiddelden of ongelijke predictor-criterium relaties voor autochtonen en allochtonen. Differentiële normen kunnen ook worden gebruikt om voor maatschappelijke achterstand te compenseren. Zo zou kunnen worden besloten om bij een toets of examen de zak-slaag grens per culturele groep te variëren of om een vooraf vastgesteld percentage allochtonen van het basisonderwijs naar het middelbaar onderwijs door te laten stromen, ongeacht het gemiddelde niveau van deze groep. Petersen en Novick (1976) hebben de psychometrische aspecten van verscheidene selectiemodellen beschreven. In de praktijk zijn differentiële beoordelingen vaak onderdeel van meer omvattende programma's zoals (in de Nederlandse literatuur) 'positieve discriminatie' en (in de Angelsaksische literatuur) 'equal opportunity' en 'affirmative action'.

Een voorbeeld van een uitgewerkt systeem met differentiële normen is ontwikkeld door Mercer (1979, 1984; cf. Cronbach, 1984, p. 209-214). Uitgangspunt is de testprestatie op de WISC-R, aangevuld met informatie over de sociaal-culturele achtergrond van het kind (onder andere gezinsgrootte, (on)volledigheid van het gezin en integratie ervan binnen de Amerikaanse samenleving). Het standaard intelligentiequotiënt wordt hier 'School Functioning Level' genoemd; deze naamgeving verwijst nadrukkelijk naar de relatie tussen school en IQ. Door de subtestscores te vergelijken met prestaties van kinderen met een vergelijkbare sociaal-culturele achtergrond – Mercer heeft een normeringsstudie uitgevoerd bij drie groepen: blanken, zwarten en 'Hispanics' – wordt een zogenaamd 'Estimated Learning Potential' verkregen. Met Mercers scoringsmethoden scoren zwarten gemiddeld 11 IQ-punten en Hispanics gemiddeld 7 IQ-punten hoger dan met Wechslers conventionele scoringsmethode. Mercers werk is sterk bekritiseerd, met name vanwege de implicaties ervan. Zo stelt Cronbach (1984) dat niet aangetoond is dat – zelfs 'with appropriate nurturance' (p. 211) – een zwart kind met een door Mercers procedure verhoogde leerpotentieelscore op school even goed zal presteren als een blank kind met een identieke score. Naast deze praktische implicatie is er een theoretische. Mercer kan niet aannemelijk maken dat haar scorecorrecties inhoudelijk adequaat zijn. Hebben een blank en een zwart kind met elk een Estimated Learning Potential van 100 een gelijk leerpotentieel? Daarvoor wordt geen evidentie naar voren gebracht. Het enige wat we weten is dat beide kinderen in hun eigen culturele groep een zelfde relatieve status hebben. Dit betekent dat een zelfde proportie van kinderen uit de eigen culturele groep een hogere of lagere score heeft.

In het Nederlandse onderwijs wordt nogal eens op enigszins intuïtieve wijze met differentiële beoordelingsmodellen gewerkt. Zo bleek uit een onderzoek van De Jong (1987) dat autochtone leerlingen in de praktijk gemiddeld negen punten hoger moesten scoren op de GALO, een toets aan het eind van het basisonderwijs, om hetzelfde schoolkeuze-advies te krijgen dan allochtone leerlingen. In hoeverre deze norm van positieve discriminatie op de lange termijn ook daadwerkelijk een positief effect heeft, is vooralsnog onduidelijk (Driessen, 1991). Zo zijn er bijvoorbeeld aanwijzingen dat bij 'overgeadviseerde' leerlingen beduidend meer sprake is van schooluitval en zitten blijven (De Jong, 1987; Driessen, 1991). Ook in eigen onderzoek zien we voorbeelden van positieve discriminatie. Noch Willemse (1989) noch Van de Rijt (1990) vonden noemenswaardige verschillen tussen allochtone en autochtone leerlingen op rapportcijfers, terwijl op de CITO-toetsen wel significante groepsverschillen werden aangetroffen. Het is denkbaar dat docenten van multiculturele klassen in het toekennen van rapportpunten meer zijn gericht op de individuele vorderingen van de allochtone leerlingen dan op het aangeven van een relatieve positie van deze leerlingen in de groep.

Een derde mogelijkheid bestaat uit het, via statistische of linguïstische procedures, *verhogen van de bruikbaarheid* van instrumenten. Deze traditie is in de Angelsaksische literatuur bekend als *'item bias'* en *'differential item functioning'* en in Nederland als *'vraagpartijdigheid'*. Deze aanpak is meestal specifieker dan de twee vorige. Storende factoren worden nu niet zozeer gezocht op test, maar op itemniveau. Bij het gebruik van testadaptaties en differentiële normen stond de bruikbaarheid van een instrument als zodanig ter discussie; in onderzoek naar vraagpartijdigheid wordt de bruikbaarheid van de items onderzocht. Nadat een test is afgenomen in twee of meer culturele groepen, worden de items afzonderlijk op hun bruikbaarheid bekeken. Dit kan op linguïstische (cf. De Jong & Vallen, 1989) of psychometrische wijze (cf. Kok, 1988). Een recent voorbeeld van een linguïstische analyse is te vinden in het werk van de Testscreeningscommissie (Hofstee, 1990; Hofstee et al., 1990). Deze beoordeelde de inhoud van twintig regelmatig gebruikte tests op hun toepasbaarheid bij allochtonen. De conclusie was negatief.

Hoewel in geen van de tests een expliciet racistische inhoud werd aangetroffen, was etnocentriciteit in met name verbale tests en vragenlijsten 'een bijna universeel verschijnsel' (Hofstee, 1990, p. 291). Als gevolg hiervan acht de commissie de toepasbaarheid van deze tests bij allochtonen sterk beperkt.

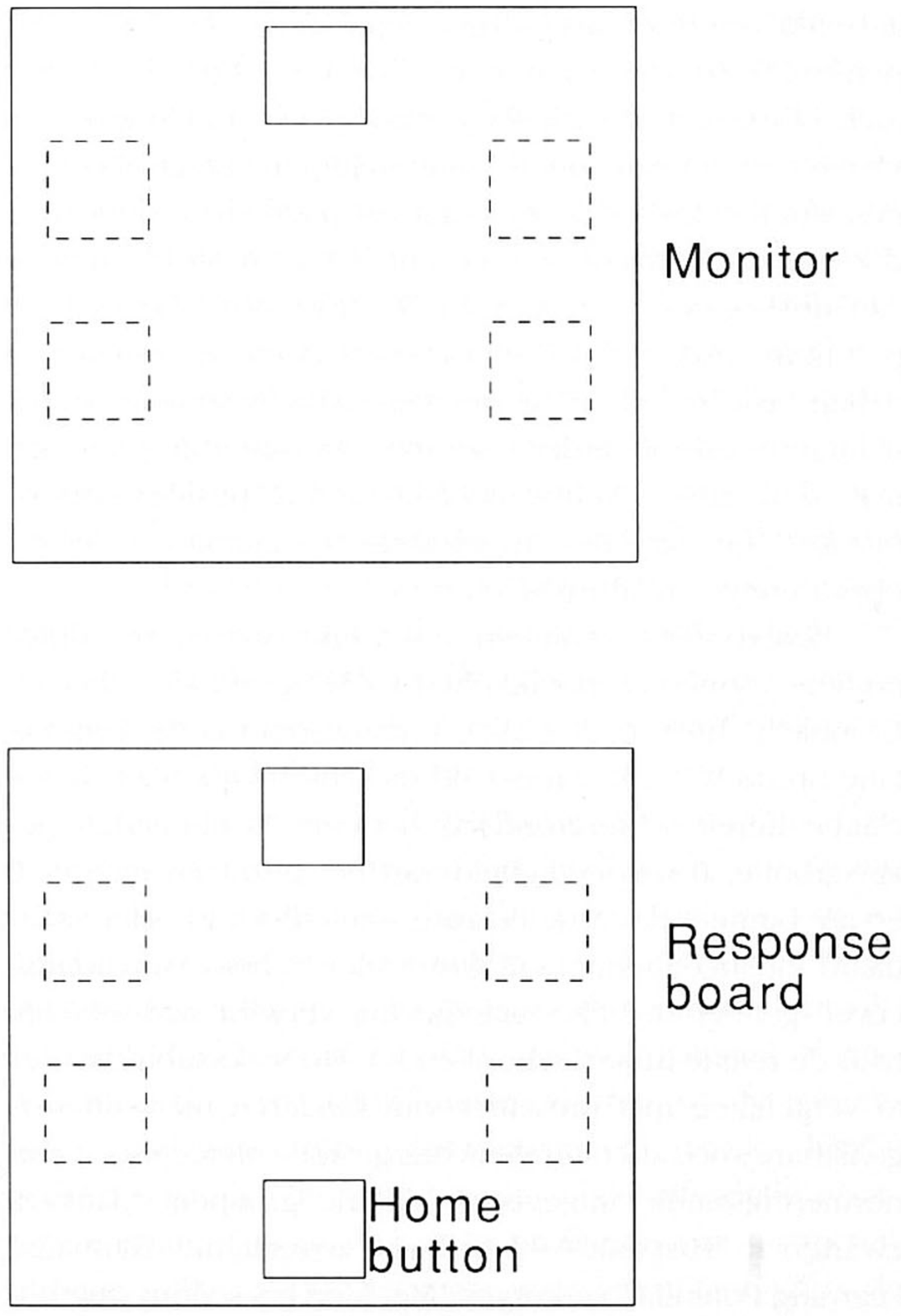
De psychometrische analyse van itembias heeft de laatste decennia een grote vlucht genomen. Een wijd scala aan technieken is ontwikkeld (een overzicht is te vinden in Berk, 1982). Behoudens een enkele uitzondering is in ons land weinig onderzoek naar itembias verricht (Hofstee, 1990). De meest uitgebreide studie is uitgevoerd door Kok (1988). Hij heeft de resultaten van autochtone, Turkse en Marokkaanse leerlingen op de CITO-Eindtoets 1982 geanalyseerd (cf. Van Esch, 1983). Taalkenmerken van de items bleken de belangrijkste bron van bias; items met 'veel woorden in het goede alternatief', 'veel woorden in het foute alternatief' en 'veel woorden in de opgave' vertoonden een bias tegen de Turkse en Marokkaanse leerlingen. Testconstructeurs die zoveel mogelijk bias proberen te vermijden, dienen derhalve zo eenvoudig mogelijk Nederlands te gebruiken; lange zinnen, moeilijke woorden en ingewikkelde zinsconstructies zijn uit den boze.

Een onderzoek waarin gebruik is gemaakt van zowel psychometrische als linguïstische analyses, is beschreven door Coenen en Vallen (1991). Een van de doelstellingen van dit project is na te gaan of een wijziging van in statistisch opzicht partijdige items andere testprestaties te zien geeft. Uitgangspunt hierbij vormt de door Uiterwijk (1990) op de Eindtoets 1987 uitgevoerde (psychometrische) Mantel-Haenszel procedure. In deze procedure worden de allochtone en autochtone proefpersonen opgedeeld naar totaalscore. De opgaven met een grote statistische itembias werden gewijzigd; de bronnen van partijdigheid werden zoveel mogelijk verwijderd. De resultaten op zowel de oorspronkelijke als de gewijzigde items van autochtone en allochtone leerlingen uit groep 8 van het basisonderwijs die vergelijkbaar waren in onder andere taalvaardigheid en score op de Eindtoets, werden bestudeerd. Allochtone leerlingen bleken hoger te scoren op de gewijzigde items dan op de oorspronkelijke items. Voor de autochtone jaargenoten bestond er nauwelijks verschil tussen beide versies. De auteurs concludeerden dat met name woordgebruik, impliciete zins- en tekstverbanden en itemcomplexiteit bronnen van itembias voor allochtonen zijn.

Nieuwe tests

Ten slotte is het mogelijk om *nieuwe instrumenten* te ontwikkelen, die problemen van conventionele tests zoals het kapitaliseren op kennis van de Nederlandse taal en cultuur zoveel mogelijk ondervangen. Zo zijn er zogenaamde 'leerpotentieeltests' ontwikkeld (onder andere Lutje Spelberg, 1987). Op het gebied van etnische minderheden is hier aandacht aan besteed door Hamers, Van Luit en Hessels (1989; Hessels & Hamers, 1993). Hun 'Leertest voor Etnische Minderheden' (LEM) is gebaseerd op Vygotsky's concept van de zone van de naaste ontwikkeling. Een onderscheid wordt gemaakt tussen wat een kind zelfstandig kan en wat het na hulp van een volwassene kan. De test bestaat uit vijf subtests waarin met name een beroep op inductief redeneren en geheugen gedaan wordt. Er wordt veel met figurale stimuli gewerkt. Aan de afname van elke subtest gaan een nonverbale testinstructie en een oefenfase vooraf waarin hulp wordt geboden op het moment dat het kind fouten maakt. Er is een normeringsonderzoek uitgevoerd onder 419 Turkse en Marokkaanse kinderen in de leeftijd van 5;4 tot en met 7;9 jaar. De test bleek een hoge betrouwbaarheid te hebben, van .88 tot .91 voor de Turkse groep en van .90 tot .92 voor de Marokkaanse groep.

De verschillen in gemiddelde leerpotentieelscore tussen allochtonen en autochtonen zijn wel kleiner dan op de RAKIT maar zijn in absolute zin verre van klein (ongeveer één standaarddeviatie). Een dergelijk groot verschil in leerpotentieel tussen deze groepen lijkt onwaarschijnlijk. Dit betekent dat men voorzichtig moet zijn met cross-culturele scorevergelijkingen op de LEM, maar het impliceert niet dat de test onbruikbaar zou zijn in een multiculturele context. Integendeel, Hessels en Hamers (1993) tonen aan dat bij Turkse en Marokkaanse leerlingen de LEM een waardevolle bijdrage kan vormen op bestaande intelligentietests; ook bij kinderen die slechts kort in Nederland zijn en de Nederlandse taal nog niet goed beheersen en

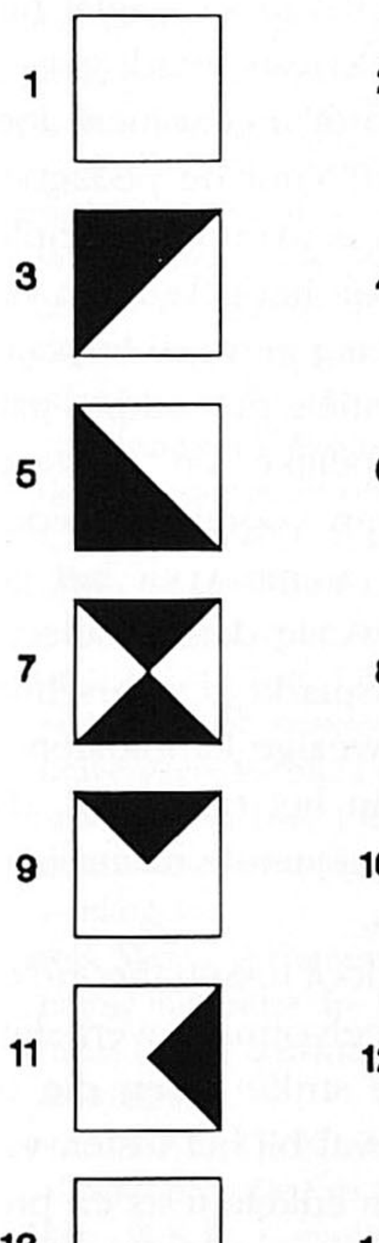


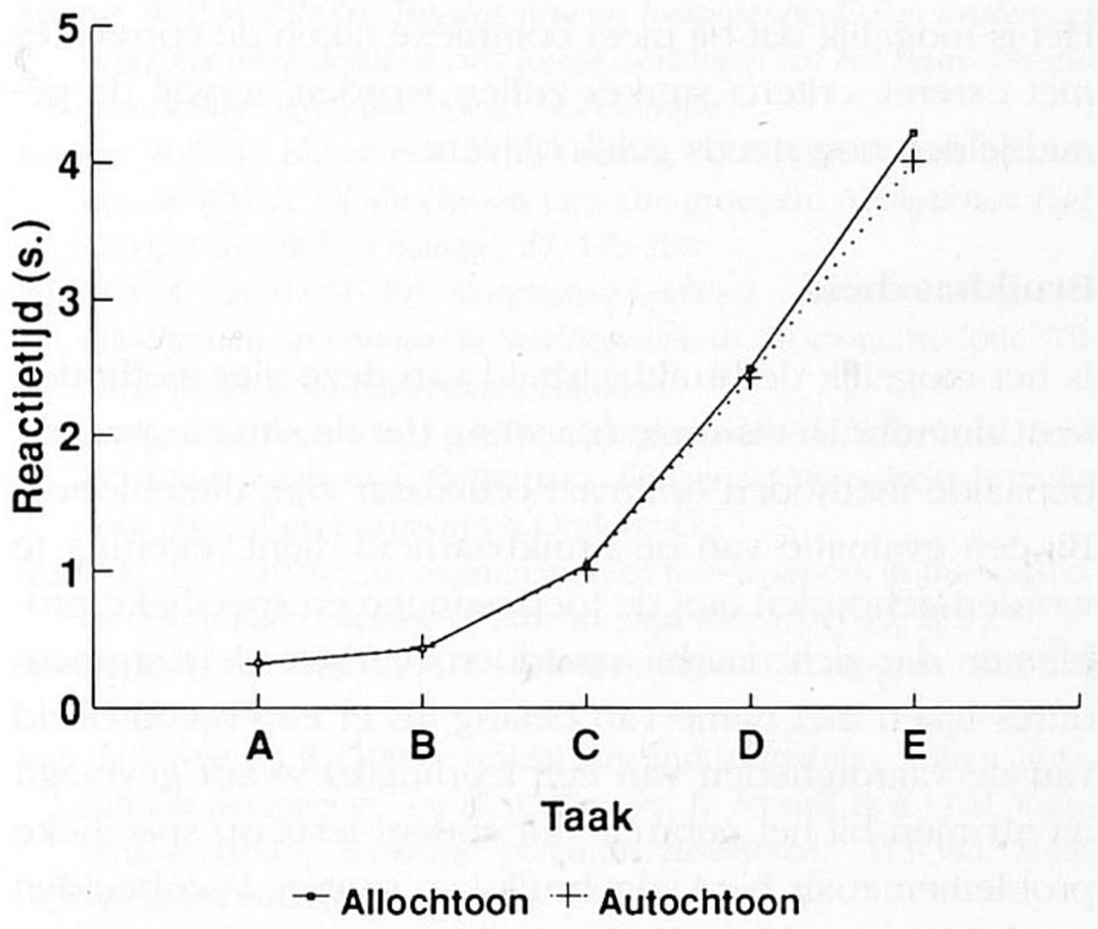
Figuur 3. Schematische presentatie van de monitor en het knoppenpaneel (de gestippelde vierkanten worden niet in de eerste taak gebruikt)

derhalve nog nauwelijks met conventionele instrumenten te testen zijn, kan het leerpotentieel middels de LEM op betrouwbare wijze worden gemeten. De predictieve validiteit van de LEM en de RAKIT zijn vergeleken (Hessels & Hamers, 1993). Criteriummaten waren scores op CITO-toetsen voor technisch lezen, begrijpend lezen, spelling en rekenen. Er waren geen verschillen in correlaties van de LEM en de RAKIT met de criteriummaten. Hessels en Hamers (1993) concludeerden dat de predictieve validiteit van beide tests gelijk zijn.

In ons eigen onderzoek werken we niet met paper-and-pencil tests maar met eenvoudige mentale taken waarbij het met name gaat om de snelheid van reageren (cf. Van de Vijver, 1993). Het belangrijkste doel van deze overgang is het verkleinen van de invloed van taal op testprestaties. De proefleid(st)er begint met steeds met een paar items voor te doen bij de testinstructie. De leerling(e) zit achter een computerscherm met voor zich een knoppenpaneel waarop afhankelijk van de taak twee tot zes knoppen zichtbaar zijn (zie Figuur 3). Er is een testbatterij ontwikkeld van vijf taken, die oplopen in cognitieve complexiteit.

De eerste taak is een enkelvoudige reactietijdtaak. Op het 'response board' zijn twee knoppen zichtbaar (de niet gestippelde knoppen op het paneel van Figuur 3), een 'home button' en een 'responsknop'. Na een auditieve waarschuwingsstimulus verschijnen de randen van een vierkant boven in het scherm. De leerling(e) moet dan de home button indrukken. Na enige seconden wordt het vierkant zwart. De leerling(e) moet dan zo snel mogelijk de bovenste knop indrukken. De taak bestaat, evenals alle overige taken, uit 20 trials. De andere vier taken zijn meervoudige reactietijdtaken; bij deze taken worden alle knoppen van het knoppenpaneel gebruikt. Op het scherm verschijnen vijf vierkanten (zie Figuur 3). Na enige tijd wordt één van de vierkanten zwart. Zodra dit gebeurt, moet de leerling(e) de ermee corresponderende knop indrukken. In de derde taak verschijnen vijf figuren op het scherm, waarvan er vier identiek zijn (bijvoorbeeld viermaal de vierde figuur van Figuur 4). Daarnaast verschijnt in iedere trial de negende figuur in een van de vierkanten. De leerling(e) dient zo snel mogelijk de knop indrukken die bij deze laatste figuur hoort. In de vierde taak worden twee paren van identieke figuren getoond (bijvoorbeeld tweemaal de vierde en tweemaal de zesde figuur) en één figuur verschijnt slechts eenmaal. Het is de bedoeling dat de leerling(e) zo snel mogelijk de knop van deze figuur indrukt. In de laatste test wordt het idee van 'complementariteit' geïntroduceerd. Twee figuren zijn complementair als deze bij samenvoeging precies één wit en één zwart vierkant vormen; Figuur 4 bestaat uit acht rijen van complementaire figuren (bijvoorbeeld figuur 3 en 4). In deze taak verschijnen twee complementaire paren figuren op het scherm, terwijl van één van de figuren geen complement aanwezig is. De leerling(e) wordt gevraagd de knop van deze laatste figuur zo snel mogelijk in te drukken.





Figuur 5. Gemiddelde reactietijden per taak en per culturele groep

Willemse (1989; Van de Vijver & Willemse, 1991) heeft de taken afgenomen bij 47 allochtone en 59 autochtone leerlingen van groep 8, in beide culturele groepen ongeveer even veel jongens als meisjes. De resultaten zijn weergegeven in Figuur 5. Hieruit blijkt dat beide groepen ongeveer hetzelfde prestatieniveau hadden. Op geen van de taken bleek een significant verschil te bestaan tussen de prestaties van allochtonen en autochtonen. Ook in Van de Rijts (1990) onderzoek bleken de prestatieverschillen op deze taken tussen beide groepen verwaarloosbaar.

De vraag rijst vervolgens of de prestaties op de taken van onze batterij aan schoolprestaties en IQ gerelateerd zijn. Dit bleek het geval in beide culturele groepen: naarmate de taken complexer werden, steeg de correlatie met schoolprestaties (rapportpunten en CITO-Entree scores). De eenvoudigste taken vertoonden geen verband met schoolprestaties. Voor de meest complexe taken waren de correlaties significant in beide groepen en schommelden rond -0.30. (De correlatie is negatief; dit wil zeggen dat

sneller reagerende proefpersonen gemiddeld hogere schoolprestaties halen.) Verder bleken de correlaties van de complexe taken met het performale IQ sterker dan met het verbale IQ (resp. -0.50 en -.030). In de literatuur is veel onderzoek gerapporteerd naar zogenaamde 'elementary cognitive operations'. Dit zijn taken waarin eenvoudige cognitieve operaties zo snel mogelijk moeten worden uitgevoerd. De eerste twee van onze batterij zijn bekende voorbeelden van deze taken. Vooral in Angelsaksisch onderzoek zijn de relaties van deze taken met schoolprestaties en scores op intelligentietests bestudeerd. In een overzicht van deze studies vond Hunt (1987) een gemiddeld verband van -0.30. Onze eigen resultaten zijn dus in lijn met de literatuur.

De resultaten van beide onderzoekingen waren interessant: de testbatterij vertoonde geen verschillen in gemiddelde score tussen de culturele groepen terwijl de prestaties erop verband hielden met rapportpunten en IQ in zowel de groep autochtonen als de groep allochtonen. Het is mogelijk dat bij meer complexe taken de correlaties met externe criteria sterker zullen worden, terwijl de gemiddelden nog steeds gelijk blijven.

Bruikbaarheid

Is het mogelijk de bruikbaarheid van deze vier methoden te evalueren? Het is onze opvatting dat de situaties waarin bepaalde methoden optimaal bruikbaar zijn, uiteenlopen. Bij een evaluatie van de bruikbaarheid dient rekening te worden gehouden met de toepassingen en specifieke problemen die zich daarbij voordoen. Verbeterde testprocedures lijken met name van belang als er een breed beeld van de vaardigheden van een leerling(e) wordt gevraagd en als men bij het gebruik van andere tests op specifieke problemen zoals het taalgebruik kan stuiten. Voorbeelden zijn het testen van jonge kinderen of van kinderen die in Nederland voor het eerst aan het onderwijs gaan deelnemen. Doorstromings- en schoolkeuze-adviezen vormen een andere toepassing. Als de studievoortgang van een leerling(e) te wensen overlaat, kunnen dergelijke instrumenten inzicht geven in de vraag of intellectuele factoren hiervoor verantwoordelijk kunnen worden geacht.

De eerst genoemde remedie, een gedetailleerd onderzoek van bestaande instrumenten, zoals door Resing, Bleichrodt en Drenth (1986) uitgevoerd, is vooral nuttig bij instrumenten die vaak worden toegepast en voor de Nederlandse populatie zijn genormeerd, zoals intelligentietests. Het geeft de gebruiker aanwijzingen over de mogelijkheden en beperkingen van deze tests in een multiculturele samenleving. Een probleem van deze benadering is dat genoemde aanwijzingen moeilijk kwantificeerbaar of in een diagnostisch advies te verwerken zijn; het 'vertalen' van de gesignaleerde tekortkomingen naar concrete remedies bij een specifieke cliënt is moeilijk. In ons eigen onderzoek bleek een Chinese leerling een verbaal IQ van 52 en een performaal IQ van 127 te hebben. Wat moet de pedagoog of psycholoog dan concluderen? De voor de hand liggende conclusie, dat de betrokken test, de WISC-R, voor deze leerling in de huidige omstandigheden geen goed instrument is, zegt weinig over de vaardigheden van het kind. Taalvaardigheid zal in het algemeen een forse invloed op de testprestaties hebben en het is op zijn minst lastig om dit adequaat in de eindconclusie te verwerken. Nieuwe benaderingen zoals de Leertest voor Etnische Minderheden (Hamers, Hessels & Van Luit, 1991) en de computergestuurde reactietijdtests lijden minder aan dit euvel. Zij kunnen een goed inzicht geven in elementaire cognitieve vaardigheden en in de intellectuele capaciteiten van een leerling(e).

In termen van de predictieve validiteit is de waarde van deze tests veelal lager dan die van conventionele paper-and-pencil tests. In ons eigen onderzoek vinden we dat de CITO-Entree scores een sterkere correlatie vertonen met schoolprestaties (rapportpunten) dan onze eigen computergestuurde tests. Dit illustreert een algemene eigenschap van dergelijke tests voor intellectuele capaciteiten: in vergelijking met schoolvorderingentoetsen hebben capaciteitentests een kleiner vermogen om toekomstige schoolprestaties te voorspellen (cf. bijvoorbeeld Altink, 1988). Schoolvorderingentoetsen stemmen qua inhoud nu eenmaal meer overeen met schoolprestaties. Het is derhalve niet realistisch te veronderstellen dat het ontwikkelen van tests die meer inzicht geven in de intellectuele vaardigheden van allochtone leerlingen ook zal leiden tot een verbeterde predictie van hun studiesucces. Nieuwe testprocedures vertonen wel een statistisch significant verband met schoolprestaties en zijn als zodanig bruikbaar in het diagnostisch proces, maar dit verband zal veelal niet sterker zijn dan wat met conventionele tests wordt gevonden. Een wat afwijkende positie wordt ingenomen door procedures met differentiële normen. Voor de pedagoog en psycholoog zijn deze procedures eenvoudig te implementeren; in essentie komt het neer op het gebruiken van *verschillende* normtabellen en zak-slaag grenzen in plaats van één. Aan het gebruik van differentiële procedures gaat echter een belangrijke maatschappelijke en politieke discussie over de wenselijkheid ervan vooraf. In Nederland lijkt in onderwijsselectie meer animo voor het gebruik van differentiële normen aanwezig dan in selectie voor de arbeidsmarkt. Op de arbeidsmarkt is waarschijnlijk het motief de beste van de aanwezige kandidaten te willen recruteren belangrijker dan in het onderwijs, dat eerder als middel wordt gezien om bestaande maatschappelijke tegenstellingen te veranderen.

Bij diagnostiek gaapt vaak een kloof tussen theorie en praktijk; de kloof tussen de rijke, veelvormige werkelijkheid die de diagnosticus ziet en de strikte eisen die de methodoloog stelt. Dit is ook het geval bij het testen van allochtonen. In het voorafgaande zijn enkele tests en procedures besproken die de eisen van de methodoloog kunnen doorstaan en die bruikbaar zijn voor de praktizerende diagnosticus. Het zou echter naïef zijn te veronderstellen dat daarmee alle problemen rondom het testen van allochtone leerlingen zijn opgelost. Nieuwe, verbeterde testprocedures vormen geen middel tegen alle kwalen. Vernieuwde procedures maken conventionele tests niet over-

bodig. Beide soorten tests lijken eerder onderling uiteenlopende toepassingen te hebben. Conventionele tests kunnen soms beter toekomstig schoolsucces voorspellen, terwijl vernieuwde procedures een beter inzicht geven in de capaciteiten van allochtone kinderen, een belangrijk gegeven voor het onderwijsproces.

Conventionele en vernieuwde procedures zullen daarom veelal *naast elkaar* worden gebruikt. Als de resultaten op beide soorten tests overeenstemmen, geeft dit evidentie dat de instrumenten waarschijnlijk goed bruikbaar zijn bij de betrokken leerling(e). Als de resultaten uiteenlopen, rijst vervolgens de belangrijke vraag hoe deze discrepantie moet worden verklaard. Bij de beoordeling hiervan is het van belang te realiseren dat conventionele procedures gevoeliger zijn voor kennis van de Nederlandse taal en cultuur, terwijl vernieuwde procedures minder door de culturele achtergrond ingekleurd zijn. Als de diagnosticus een adequate verklaring kan vinden voor de discrepantie, is daarmee zijn of haar taak beëindigd. Vanuit onderwijskundig perspectief vormt zo'n observatie echter een begin. Het zou aanleiding kunnen zijn voor allerlei onderwijskundige maatregelen zoals *remedial teaching*, extra aandacht in de klas, of andere speciale opdrachten die tot doel hebben de discrepantie tussen aanwezige en gerealiseerde vermogens wat kleiner te maken.

Literatuur

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Altink, W.M.M. (1988). *Selectie voor hoger onderwijs in ontwikkelingslanden*. Utrecht: Elinkwijk.

Berk, R.A. (Red.) (1982). *Handbook of methods for detecting item bias*. Baltimore: John Hopkins University Press.

Coenen, M. & Vallen, T. (1991), Itembias in de Eindtoets basisonderwijs. *Pedagogische Studiën, 68*, 16-26.

Cronbach, L.J. (1984). *Essentials of psychological testing* (4th ed.), New York: Harper & Row.

Driessen, G. (1991), Discrepanties tussen toetsresultaten en doorstroomniveaus. Positieve discriminatie bij de overgang basisonderwijs-voortgezet onderwijs? *Pedagogische Studiën, 68*, 27-35.

Esch, W. van (1983). *Toetsprestaties en doorstroomadviezen van allochtone leerlingen in de zesde klas van lagere scholen*. Nijmegen: Instituut voor Toegepaste Sociologie.

Hamers, J.H.M., Luit, J.E.H. van & Hessels, M.G.P. (1989). *Leertest voor etnische minderheden. Eindrapport deel A*. Utrecht: Rijksuniversiteit, Faculteit der Sociale Wetenschappen.

Hamers, J.H.M., Luit, J.E.H. van & Kessels, M.G.P. (1991). *Leertest voor etnische minderheden. Test en handleiding*. Lisse; Swets & Zeitlinger.

Hessels, M.P.G. & Hamers, J.H.M. (1993). A learning potential test for ethnic minorities. In: J.H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (red.), *Learning potential assessment*, 285-311. Lisse: Swets & Zeitlinger.

Hofstee, W.K.B. (1990), Toepasbaarheid van psychologische tests bij allochtonen. *De Psycholoog, 25*, 291-294.

Hofstee, W.K.B., Campbell, W.H., Eppink, A., Evers, A., Joe, R.C., Koppel, J.M.H. van de, Zweers, H., Choenni, C.E.S. & Zwan, T.J. van de (1990). *Toepasbaarheid van psychologische tests bij allochtonen*. Utrecht: LBR reeks nr. 11.

Hunt, E.B. (1987). The next word on verbal ability (1). In: P.A. Vernon (red.), *Speed of information-processing and intelligence*, 347-392. Norwoord, NJ: Ablex.

Irvine, S.H. (1979). The place of factor analysis in cross-cultural methodology and its contribution to cognitive theory. In: L. Eckensberger, W. Lonner & Y.H. Poortinga (red.), *Cross-cultural contributions to psychology*. 300-341. Lisse: Swets & Zeitlinger.

Jong, M.J. de (1987). *Herkomst en kansen. Allochtone en autochtone leerlingen tijdens de overgang van basis naar voortgezet onderwijs*. Lisse: Swets & Zeitlinger.

Jong, M. de & Vallen, T. (1989), Linguïstische en culturele bronnen van itembias in de Eindtoets Basisonderwijs van leerlingen uit etnische minderheidsgroepen. *Pedagogische Studiën, 66*, 390-402.

Kok, F.G. (1988). *Vraagpartijdigheid. Methodologische verkenningen*. Amsterdam: De Amstel.

Mercer, J.R. (1979). *System of multicultural pluralistic assessment (SOMPA): Technical manual*. New York: The Psychological Corporation.

Mercer, J.R. (1984). What is a racially and culturally nondiscriminatory test? A sociological and pluralistic perspective. In: C.R. Reynolds & R.T. Brown (red.), *Perspectives on bias in mental testing*, 293-356. New York: Plenum.

Maussen, L.H.M. (1988). *Otis test M. Test voor verbale intelligentie*. Nijmegen.

Petersen, N.S. & Novick, M.R. (1976), An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3-29.

Resing, W.C.M. (1990). *Intelligentie en leerpotentieel: Een onderzoek naar het leerpotentieel van jonge leerlingen uit het basis- en speciaal onderwijs*. Lisse: Swets & Zeitlinger.

Resing, W.C.M., Bleichrodt, N. & Drenth, P.J.D. (1986), Het gebruik van de RAKIT bij allochtoon etnische groepen. *Nederlands Tijdschrift voor de Psychologie, 41*, 179-188.

Rijt, B.A.M. van de (1990). *Reactiesnelheidstest. Een aanvulling voor allochtonen op bestaande intelligentietests*. Doctoraalscriptie. Tilburg: Katholieke Universiteit Brabant.

Roelandt, T. & Veenman, J. (1988). *Minderheden in Nederland. Positie in het onderwijs*. Rotterdam: Erasmus Universiteit, Instituut voor Sociaal en Economisch Onderzoek.

Sarnacki, R.E. (1979), An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49*, 252-279.

Spelberg, H.C. Lutje (1987). *Grenzentesten*. Groningen: Stichting Kinderstudies.

Van de Vijver, F.J.R. (1993). Trainability and learnability from a cross-cultural perspective. In: H.M. Hamers, K. Sijtsma & A.J.J.M. Ruijssenaars (red.), *Learning potential assessment*. 313-340. Lisse; Swets & Zeitlinger.

Van de Vijver, F.J.R. & Poortinga, Y.H. (1992), Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17-24.

Van de Vijver, F.J.R. & Willemse, G.R.W.M. (1991). Are reaction time tasks better suited for ethnic minorities than paper-and-pencil tests? In: N. Bleichrodt & P.J.D. Drenth (red.), *Contemporary issues in cross-cultural psychology*. 450-464. Lisse: Swets & Zeitlinger.

Willemse, G.R.W.M. (1989). *Keuze-reactietijdtaken in multi-culturele context*. Doctoraal-scriptie. Tilburg: Katholieke Universiteit Brabant.

Drs. A.J.R. van de Vijver is als universitair docent werkzaam bij de faculteit Sociale Wetenschappen van de Katholieke Universiteit Brabant.
Mw.drs. G.R.W.M. Willemse en mw.drs. B.A.M. van de Rijt zijn beiden als Assistent in Opleiding werkzaam bij de vakgroep Pedagogiek van de Rijksuniversiteit Utrecht.